Predicting Term-Relevance from Brain Signals

Manuel J. A. Eugster[2,*], Tuukka Ruotsalo[1,*], Michiel M. Spapé[1,*],
Ilkka Kosunen[3], Oswald Barral[3], Niklas Ravaja[1,4], Giulio Jacucci[1,3], Samuel Kaski[2,3]
Helsinki Institute for Information Technology HIIT
[1]Aalto University, PO Box 15600, 00076, Finland
[2]Aalto University, Department of Information and Computer Science, PO Box 15400, 00076, Finland
[3]University of Helsinki, Department of Computer Science, PO Box 68, 00014, Finland
[4]University of Helsinki, Department of Social Research, PO Box 54, 00014, Finland
first.last@hiit.fi

ABSTRACT

Term-Relevance Prediction from Brain Signals (TRPB) is proposed to automatically detect relevance of text information directly from brain signals. An experiment with forty participants was conducted to record neural activity of participants while providing relevance judgments to text stimuli for a given topic. High-precision scientific equipment was used to quantify neural activity across 32 electroencephalography (EEG) channels. A classifier based on a multi-view EEG feature representation showed improvement up to 17% in relevance prediction based on brain signals alone. Relevance was also associated with brain activity with significant changes in certain brain areas. Consequently, TRPB is based on changes identified in specific brain areas and does not require user-specific training or calibration. Hence, relevance predictions can be conducted for unseen content and unseen participants. As an application of TRPB we demonstrate a high-precision variant of the classifier that constructs sets of relevant terms for a given unknown topic of interest. Our research shows that detecting relevance from brain signals is possible and allows the acquisition of relevance judgments without a need to observe any other user interaction. This suggests that TRPB could be used in combination or as an alternative for conventional implicit feedback signals, such as dwell time or click-through activity.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Relevance feedback; H.5.2 [**User Interfaces**]: Evaluation/Methodology

Keywords

Brain Signals; Relevance Prediction; EEG

1. INTRODUCTION

Relevance prediction is a central challenge of Information Retrieval (IR) research as it determines the information presented to the user. In this paper, Term-Relevance Prediction from Brain Signals (TRPB) is proposed to automatically detect relevance of information directly from brain signals. Research has begun to build an understanding of the neural activity associated with relevance detection [25], but 1) the exact brain areas associated with relevance judgments remain unknown and 2) the methods to non-intrusively quantify the neuronal activity can be very noisy [19]. Nevertheless, the advantages of using brain signals to predict relevance are 1) that recording brain signals does not require any explicit user interaction, and 2) the signals can be captured with high throughput. That is, TRPB only requires participants to examine text stimuli (e.g., text displayed on the screen), and the signals can be continuously recorded. Consequently, TRPB can be used in combination with or even as an alternative for conventional implicit feedback signals, such as dwell time or click-through activity, that have been shown to be unreliable and are tightly connected to conventional user interfaces and explicit user interaction [13].

We use multi-view machine learning to solve the prediction problem and to cope with the uncertainties related to active brain areas and signal noise. The multiple kernel learning approach [9, 37] allows us to incorporate the traditionally complementary approaches to electroencephalography (EEG)—time-based (event-related potentials, ERPs) and frequency-based—simultaneously in order to maximize TRPB predictive power. We demonstrate that predicting relevance judgments is possible from brain signals alone, i.e., without any explicit user interaction or brain-computer interface (BCI) training. The high throughput of brain signals enables the capture of a huge number of relevance judgments for text stimuli in a relatively short time. In practical applications it is possible to sacrifice recall and target high precision to take advantage of the high throughput. We demonstrate a high-precision variant of a relevance predictor, which is able to detect terms of the users' given topics of interest.

We conducted an experiment in which EEG signals of 40 participants were recorded when they judged relevance of text stimuli. The main findings of TRPB using these data are the following:

SIGIR'14, July 6–11, 2014, Gold Coast, Queensland, Australia.
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-2257-7/14/07 ...$15.00.
http://dx.doi.org/10.1145/2600428.2609594.

[*] Equal contributions; authors in alphabetical order

1. EEG signals can be used to automatically predict relevance with an improvement of 17% in accuracy.
2. Relevance prediction can be done independently of the user, without requiring user-specific calibration.
3. A high-precision variant of the classifier can automatically construct term representations of topics of interest with a precision increase of 30% while still maintaining a feasible recall.

We were also able to confirm that the learned parameters of the classifiers (kernel weights) gave pointers to significant changes in brain activity. In an extended analysis we find significant activity in brain areas that have previously been found responsible for recognition and memory recall, and for organizing, maintaining, and implementing intentions [18]. This supports TRPB from a cognitive science perspective.

The rest of the paper is organized as follows. Section 2 reviews the related work ranging from mind-reading to brain-computer interfacing and using sensory signals as relevance feedback. Section 3 describes the experiment that was conducted to record participants' neural activity. Section 4 presents the term-relevance prediction experiment, and Section 5 discusses the results of the experiments dividing them into classification results (5.1) and physiological findings (5.2). Section 6 demonstrates the effectiveness of TRPB in practice with a high-precision classifier. We conclude and discuss the future work in Section 7.

2. RELATED WORK

Our work is related to a range of research from understanding relevance as it is associated with brain activity, operationalizing these associations as sensory input for predicting relevance as in relevance feedback research, interfacing between a computer and a human brain, and detecting patterns form brain activity without explicitly looking for a pre-known activity, as in mind-reading research.

Relevance. In information retrieval, relevance is a widely operationalized concept. A huge body of research exists that attempts to make use of relevance in practical systems, such as relevance feedback for improving retrieval, or relevance judgement to produce ground truth for evaluation purposes [36]. Cognitive scientists have long been interested in mapping basic cognitive functions that are highly related to perceiving relevance (e.g., recognition and memory recall [32, 42]) and reacting to relevant stimuli (e.g. implementing intentions [18]). The wealth of knowledge from various fields underlines the fundamental complexity of relevance, which may be the reason why the question of "how does relevance happen in the brain" remains unanswered [25].

Relevance prediction. In information retrieval, relevance judgments of the presented information can be acquired from user interaction and behavior data. Previously, researchers have made use of explicit [16], implicit [13], or affective [1, 24] user signals and then used features extracted from these signals to build models that can be used to automatically predict relevance of information. Among explicit interactions, such as typing queries, implicit monitoring of user actions has been found the most practical source of user signal, since it is less intrusive for the user and does not require users to explicitly provide relevance judgments [13]. While previous research has found evidence that implicit behavioral signals, such as dwell time and click-through activity, can be predictors of users' information needs, they can be noisy and unreliable [14, 40]. Collecting evidence of click-through activity and dwell time also requires previous explicit interaction between the user and the information retrieval system as well as monitoring of the users over a substantial amount of time. Explicit interaction is thus increasingly challenged by new information access media, such as augmented reality interfaces, which can make collecting conventional implicit feedback impractical, but allow wearable sensors to be used to capture additional user signals.

Sensory signals. Recently, sensory signals have been utilized to measure human emotion and map emotional states to predict relevance [27, 28]. Such sensors can detect changes often expressed through a psychophysiology. Psychophysiology is reflected via cues, such as facial expressions, changes in the electrodermal activity [2], or variations in the skin temperature [6]. These physiological signals have provided researchers with additional sources of information not previously available, and their effectiveness has been empirically studied [1, 2]. However, the results seem to be contradictory and validated only for image or video stimuli. These media allow measuring users for extended periods of time, which in itself is known to cause emotional responses and more substantial physiological responses [24, 25].

Brain-Computer interfacing. A related research field that has made use of EEG signals to allow (non-invasive) interfaces to control computers is brain-computer interfacing (BCI). The key difference that sets BCI apart from our research is its requirement for user-specific memorizing. BCI typically involves a user-specific training step in which the user is required to memorize a motoric action, for instance, pulling the left arm. The BCI system is then trained to explicitly detect such previously established behavior [41] in the control phase. Therefore, BCI does not detect the associated, natural brain patterns related to relevance as such, but creates an additional, "artificial" pattern requiring extensive, conscious training. Given that a detection of relevance may also appear subliminal, the methods used in BCI are of limited use for the present study.

Mind-Reading. Mind-reading aims to use neuroimaging, typically functional Magnetic Resonance Imaging (fMRI), to learn specific patterns of brain activity with labeled object stimuli in order to predict each of these different labels on an instance-by-instance basis from the fMRI data [23]. For example, when humans think of an object, many different areas of the brain activate. This pattern can then be learned by a machine learning system (based on the blood-oxygen-level-dependent activity). Conversely, the conventional neuroimaging research aims at finding correlates to external regressors such as task condition with activity in specific brain areas. The present approach to term-relevance prediction is similar to mind-reading in that we likewise aim to directly associate brain patterns with users' subjective perception of stimuli. However, the prediction of relevance does not require association of specific patterns to specific objects but rather abstracting of a general pattern of brain activity in order to predict relevance for a stimulus.

Unique contributions. To our knowledge, our work is the first to predict term-relevance from brain signals for an IR scenario. What sets our research apart from the related research is the following: 1) we use brain signals captured via EEG alone without any other user signal, 2) we use text

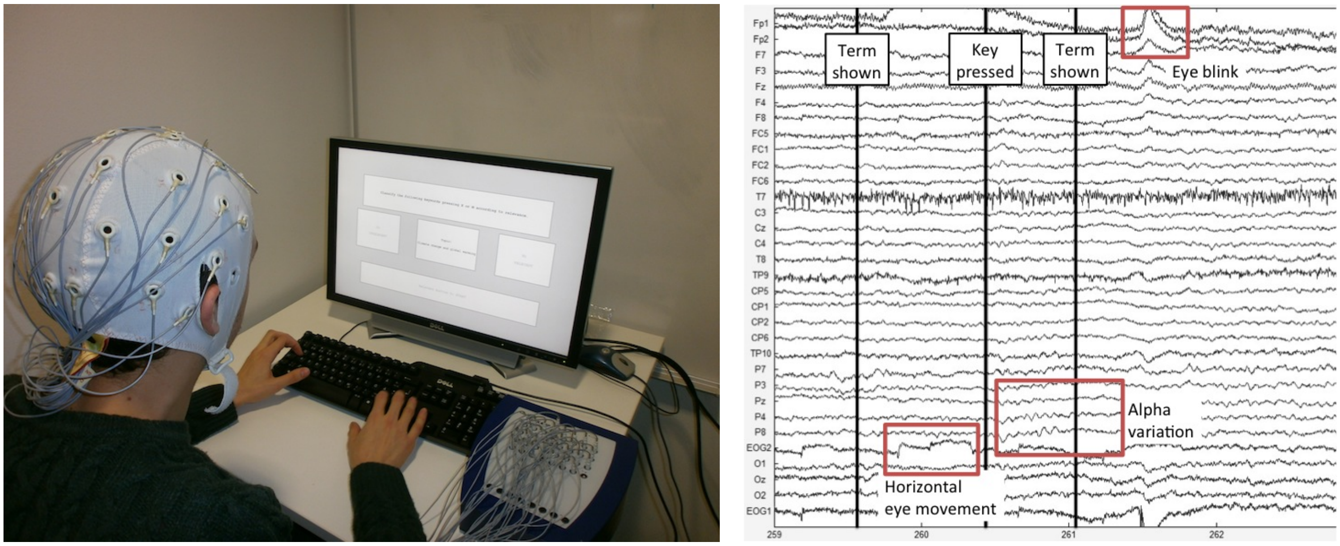

Figure 1: (a) A participant of the experiment with the full EEG sensor setup to record the raw EEG signals reading the instructions for the next task. (b) Excerpt of the participant's captured raw EEG signals showing one trial from term onset to key press to the next term onset. The x-axis shows the time in milliseconds, the y-axis the different EEG channels. The annotations show a typical pattern of an eye blink, horizontal eye movements, and variation in the alpha frequencies. An artifact correction procedure was then applied to remove, for example, eye movement related activity (see Section 3.4), before specific features were extracted (see Section 3.5).

stimuli, 3) we aim to learn brain patterns that are naturally associated with relevance judgments rather than to detect artificial, memorized patterns or pre-seen objects, 4) we aim for generalization over participants such that we can learn the underlying patterns of brain activation and use them for an unseen content and unseen participants.

3. NEURAL-ACTIVITY RECORDING EXPERIMENT

We recorded the EEG signals of participants when they assessed relevance in response to term stimuli shown on a screen. The term stimuli were associated with a predetermined topic. For example, participants were asked to judge relevance of terms "Snowmelt" and "hardware synchronization" for a topic "Climate change and global warming" (the first term is relevant, the second term is irrelevant). Terms were chosen to represent real search situations where terms are not always clearly discriminative. For example, the term "Morse code" is not clearly irrelevant for the topic "Iraq war", and participants might differ in their relevance judgments. We used a highly controlled experimental setup to avoid possibly confounding effects related to hemispheric lateralization, eye-movements, and motor activity. For example, moving the eyes to read a word or moving the arm to give the relevance judgment are visible in the brain signals and can result in a classifier that learns, for example, the effect of the arm-movement and not the effect of the relevance judgment.

3.1 Experimental Design

Each participant judged relevance for six terms in six topics, for a total of 36 trials. The terms were randomly drawn from a pool of relevant (for each topic) and irrelevant (for all topics) terms defined by experts (column "Topic" in Table 4 lists all six topics; column "Predicted top 5 relevant terms" gives an excerpt of the terms). We used a balanced setup, i.e., for each topic three relevant and three irrelevant terms were shown to the participant. We randomized the order of the topics and the terms over participants. In addition, the relevance-key assignment was balanced (right or left hand used) between blocks of 12 trials to avoid possibly confounding hemispheric effect. The presented items were balanced to be apriori 50% relevant and 50% irrelevant. This random baseline, although not entirely realistic, ensures that we measure signals and effects related to relevance judgments and not, for example, related to the well-known P300 effect in an oddball-paradigm based experiment [10]. The recoded data reflect the user's subjective relevance judgment of the items (i.e., if a participant assessed the apriori irrelevant "Morse code" relevant for the topic "Iraq war", it was recoded as relevant), as this is the user's real assignment and the corresponding effect is what we would also expect to predict from the brain signals.

3.2 Participants

Forty participants, 34 males and 6 females, took part in the study. The age of the participants ranged from 21 years to 47 years (Mean = 28.17, Median = 26.5). Most of them were post-graduate (37), and the rest were undergraduate students. Only two of the participants reported to be English native speakers, seventeen different mother tongues being reported. Nevertheless, English reading skills were overall reported as advanced (Mean = 4.55, Median = 5; on a scale from 1 to 5). Six of the participants were left-handed. Participants reported themselves as physically and mentally healthy. Participation was compensated with two movie tickets.

Table 1: The outcome of the neural-activity recording experiment: For each term presented to a participant, we collected its binary relevance judgement. Then, seven views were computed with features based on the recorded EEG signal during a certain period of time from term stimulus onset until corresponding relevance judgments. The 20 features represent the 20 most central channels F3, Fz, F4, FC5, FC1, FC2, FC6, C3, Cz, C4, CP5, CP1, CP2, CP6, P3, Pz, P4, O1, Oz, and O2.

Views	$\mathbf{v}_k$	Features
Relevance judgement view:		
Relevance		A binary relevance judgement provided by a participant for a term for a given topic
Frequency-band-based views:		
Theta	1	40 features for each frequency band: 20 features of average power over 1 second epochs before the relevance judgement; 20 features of average power over entire period, minus power of the second before term onset
Alpha	2	
Beta	3	
Gamma1	4	
Gamma2	5	
Engage	6	
Event-related-potential-based view:		
ERPs	7	80 features of average amplitude: 20 features for 80–150 ms, P1; 20 features for 150–250 ms, N1/P2; 20 features for 250–450 ms, N2 or P3a; 20 features for 450–800 ms: N4 or P3b

3.3 Procedure

At the beginning of the session the participants were briefed as to the procedure and purpose of the experiment. Then they were asked to sign an informed consent. They were furthermore informed on their right to withdraw from the experiment at any moment without any adverse consequence. The task was explained in more detail prior to the execution. The participants were instructed to rate the text stimulus either relevant or irrelevant by pressing either the M-key (using the right hand) or the X-key (using the left hand), on a QWERTY keyboard. They were instructed to provide the relevance judgement by pressing the key as soon as they were able to make the judgement. The next term was presented as soon as the participants pressed the relevance key. After the experiment, the participants were asked to fill out an online questionnaire regarding their background information.

3.4 EEG Recording and Processing

A QuickAmp (BrainProducts GmbH, Gilching, Germany) amplifier recorded EEG at a sample rate of 100 Hz. EEG was recorded from 30 Ag/AgCl scalp electrodes, positioned using EasyCap elastic caps (EasyCap GmbH, Herrsching, Germany) on equidistant electrode sites of the 10% system excluding FT9/FT10. Figure 1 shows a participant with the full EEG sensor setup. Processing of EEG was conducted in EEGLAB [31] and included re-referencing to the common average reference and filtering of the data between 1 and 80 Hz with a notch filter between 46 and 54 Hz to reduce DC interference. After that, an automatic artifact correction, based on the Efficient Independent Component Analysis algorithm [17], as implemented in the AAR toolbox [8], was carried out in order to eliminate noise and potential confounds of common artifacts such as eye movements and blinks (see Figure 1).

Visual inspection of the raw data revealed extreme noise levels for two participants (possible, for example, because of loose electrodes or a cap which does not fit exactly). These participants' data were removed from further analysis, which left us with $S = 38$ participants. In addition, we only considered judgments that conformed with the ground truth to reduce noise induced by judgments possibly done by chance when participants were not sure about their judgement.

3.5 Feature Engineering

Frequency-band-based features (FBF) and event-related-potential-based features (ERPF) were extracted from the pre-processed signals. FBFs capture the change in the signals for the whole time window when the user was shown the stimulus. ERPFs capture the changes in the signals for a specific short time window when a participant makes the relevance judgement—which can be a much shorter time window and not necessarily at the time of giving the explicit relevance judgement but, e.g., a few (milli)-seconds after the term was shown on screen.

As no consensus exists on where and how binary relevance judgments of text stimuli affect neural activity, it was not possible to focus on, e.g., one specific frequency band or brain area. Therefore, we engineered a set of different FBFs and RPRFs in order to capture all the data that are potentially related to the relevance judgement. In both cases, the EEG was time-locked to the start (i.e., term shown on screen) or end (i.e., participant gave the relevance judgement) key events in the experiment. Table 1 gives a summary of the seven views and the corresponding features.

In order to maximize the cortical activity signal and minimize muscle-related activity and other artifactual noise, we included only the 20 centrally located electrodes. To obtain features, we calculated the power of the segment of 1 second following the term onset using the fast Fourier transform and applying log-transformation to normalize the signal. From this, a baseline was subtracted by the same procedure over the 1 second prior to the term onset.

Frequency-band-based views. An essential aspect of electroencephalography (EEG) is that different types of oscillations, from the very slow theta (4–8 Hz) to the higher gamma (80 Hz), have been associated with various psychological functions. For example, alpha activity has been related to attentiveness [3], theta activity to attention [22], and alpha desynchronisation with semantic memory performance [15]. Possibly, decisions regarding relevance or irrelevance, through acts of motor imagery [21] and motor control [38], would trigger activity in the beta frequency. Finally, given previous indications of the role of gamma activity in consciousness [4], one might expect relevant search results to be particularly accessible to consciousness and thus be associated with gamma activity. Further evidence for this comes from the observation that gamma-band oscillations have been associated with attentional information processing through the salience of stimuli [12].

Furthermore, combinations of multiple frequency bands have also been shown to account for cognitive functions. For example, a combination of theta, alpha and beta bands has been found to be an index of engagement [31], which we therefore include here as another candidate. Other combinations of frequency bands were also tested from within the multi-view model as will be discussed further on.

Event-related-potential-based view. Event-related-potentials (ERPs) are brain responses resulting from specific sensory, cognitive or motor events as measured using EEG. Generally, as stimuli are sensed, the modality-specific sensory areas in the brain are activated early, appearing in the EEG as peaks with a specific topography, latency and direction (negative or positive).

A set of ERPs have been associated with cognitive functions [20]. For example, the negative, fronto-central N2 has been associated with uncertainty and cognitive control [7], which could be related to the task of information retrieval. The P3 potential occurs generally after 300 ms and is commonly separated in two sub-components called the P3a and P3b. The P3a has more frontal topography than the P3b and is associated with orientation and attention while the P3b is related to memory processing and retrieval [30]. The P3 is also one of the earliest potentials to be used for the purposes of BCI [5].

As with the frequency analyses, no apriori decision was made to exclude specific potentials. Instead, we calculated the cross-individual average of the ERP and defined the intervals, based on the literature and visual inspection, to occur at 80–150 ms (P1), 150–250 ms (N1/P2), 250–450 ms (N2 or P3a) and 450–800 ms (N4 or P3b) for all the 20 electrodes mentioned earlier, relative to the 200 ms prior to the onset of the term, thus constituting 80 features in total.

4. RELEVANCE PREDICTION FROM BRAIN SIGNALS

Given the feature representation of the collected neural-activity data, we studied:

1. How well can we predict relevance judgments on terms from the brain signals of unseen participants?

2. Which EEG views are important for the prediction?

The first question is motivated by real search situations in which no user-specific training or calibration is necessarily possible. The second question is motivated by the currently unknown brain areas associated with relevance judgments; an answer to this question allows us to draw some conclusions about the cognitive basis of the brain areas that are found to be important for relevance prediction.

To answer these questions we have devised a set of binary relevant/irrelevant classification experiments based on a multi-view learning method and a leave-one-participant-out strategy. Multi-view learning is the task of learning from two or more data sets with co-occurring observations. This concept perfectly suits our scenario, and we used it by treating the different representations of the EEG signals as different views of a relevance judgement given by a participant. Formally, each relevance observation $r_i = (y_i, \mathbf{v}_1, \ldots, \mathbf{v}_K)$ is described by the binary relevance judgment y_i and by K different feature vectors $\mathbf{v}.$ (i.e., views). For each participant s, we have N^s relevance observations, i.e., $R^s = \{r_1^s, \ldots, r_{N^s}^s\}$ is the set of relevance observations pertaining participant s. The $R = \{R^1, \ldots, R^S\}$ is the set of all relevance observations across all participants.

4.1 Multiple Kernel Learning

We use multiple kernel learning (MKL) support vector machines [37] as a multi-view learning method to learn classification models of the form

$$y = f(\mathbf{v}_1, \ldots, \mathbf{v}_K) = \sum_{k=1}^{K} \beta_k \langle \mathbf{w}_k, \Phi_k(\mathbf{v}_k) \rangle + b,$$

given a set of relevance observations $\{r_i\}_{i=1}^N$ as learning data. Here y denotes the binary relevance judgement, $\langle \cdot, \cdot \rangle$ the scalar product, $\mathbf{w}_k$ the weight vector of the observations, $\Phi_k(\mathbf{v}_k)$ the feature map of the view $\mathbf{v}_k$, β_k the kernel weights, and b the bias. The learning problem is to estimate the optimal kernel weights β_k along with $\mathbf{w}_k$ and b from the given data. Using different feature maps (and consequently different kernels) allows us to represent the fact that the different EEG signals can have different measures of similarities, and we can capture nonlinear relationships between features. The estimated kernel weights β_k can be used as an indication for the importance of the different views [37], which, consequently, allows us to draw conclusions on the importance of individual EEG signals in predicting relevance.

For the concrete estimation of the classification models, we use a Bayesian MKL algorithm with an efficient inference based on variational approximation [9]. Among other advantages, the Bayesian formulation of MKL estimates the predictive distribution of the class labels. We will later on utilize the predictive distributions when constructing the high-precision classifier (Section 6). For concrete details, especially on the actual model specification, the distributional assumptions, and the formulation of the deterministic variational approximation, we refer to [9].

4.2 Prediction Setup

Our prediction setup is based on the data of $S = 38$ participants and $K = 7$ views $\mathbf{v}_k$ with features described in Table 1. We computed models with different combinations of views. We applied a leave-one-participant-out learning strategy as follows. For each participant s we learned a classification model f_s using the other participants' data (i.e., the learning set $R^{\bar{s}} = R \setminus R^s$) with the views that were selected for the particular model (e.g., All views and Alpha+ERPs). The prediction accuracy was then computed on the participant's relevance observations, i.e., the test set R^s. The learning of the models results in estimated observation weights $\mathbf{w}_k^{\bar{s}}$, kernel weights $\beta_k^{\bar{s}}$, and the bias $b^{\bar{s}}$.

We used an automatic feature selection procedure on each view $\mathbf{v}_k$, whereby the features were ranked according to the t-statistic (computed between the relevant and irrelevant observations) and the highest ranking features were selected [35]—in our case the top ten features. Given the selected features for each view $\mathbf{v}_k$, we normalized the data and computed a Gaussian kernel with the kernel width defined as the median distance between the observations [11].

The number of relevance observations N^s varies slightly for each participant because we used only observations that conformed to the ground truth. We established balanced data by randomly drawing the learning set and the test set from the set of relevant and the set of irrelevant observations, each with the number of observations defined by the smaller set. This reassembles our original experimental design and is a simple but well-established strategy to exclude possi-

Table 2: Classification results based on all 38 participants for different sets of views. The table lists the mean classification accuracy, the p-value indicating a significant better mean classification accuracy than the random baseline, and the corresponding mean improvement. Because of our experimental design, the random baseline prediction of whether a term is relevant or irrelevant is 0.5. Bold entries denote that improvements are statistically significant at a level of $\alpha = 0.01$, p-value $< \alpha$ with correction for multiple testing.

Views	Mean accuracy	p-value	Mean improvement
All	0.5415	**0.0003**	8.30%
Individual views:			
Alpha (Al)	0.5242	0.0265	4.83%
Gamma1 (Ga1)	0.5143	0.1445	2.86%
Beta (Be)	0.5005	0.4838	0.10%
Gamma2	0.5101	0.2003	2.02%
Theta	0.5000	0.4984	0.01%
ERPs (E)	0.5312	0.0092	6.24%
Engage	0.4773	0.9673	−4.55%
Selected combined views:			
Al+Ga1	0.5429	0.0014	8.59%
Al+E	0.5475	**0.0007**	9.50%
Ga1+E	0.5528	**0.0002**	10.55%
Al+Ga1+Be	0.5369	0.0022	7.37%
Al+Ga1+E	0.5586	<**0.0001**	11.72%

ble problems of the classification method with imbalanced classes. To eliminate a possible observation sampling bias we repeated this procedure five times; i.e., for a participant s we estimated five models f_s^i $(i = 1, \ldots, 5)$ and consequently five estimations of model parameters. We report averaged results, unless otherwise noted.

5. RESULTS

In this section, we first present results from classification experiments with various combinations of EEG views. We discuss the importance and influence of the various EEG views and—encouraged by the well-known "BCI illiteracy"—we show that there exists a restricted set of participants for which we can further improve the prediction accuracy. We then show physiological findings that map the important views to effects that can be localized to certain brain areas.

5.1 Classification Performances

Table 2 summarizes the classification accuracies for different sets of views. We report the mean classification accuracy, improvement over the random baseline, and the p-value of a t-test for significance corrected for multiple testing using the Bonferroni correction. The t-test was applicable because the Shapiro-Wilk test showed no significant difference from the normal distribution. The classifiers using all seven EEG views (All) predicted relevant and irrelevant terms for an unseen participant significantly better than the random baseline, and achieved a mean improvement of 8.30%.

Importance and influence of EEG views. The estimated kernel weights β_k of all learned classification models using all seven views gave us a first indication of the importance of each EEG view. Alpha and Gamma1 have the highest weights, then Beta, then Gamma2, Theta and ERPs, and finally Engage. To study the influence of different views on the classification accuracy, we built models for each view separately. The corresponding results are shown in the middle block of Table 2. These single-view runs indicated that none of the individual views alone led to significant improvements. However, we found that Alpha and ERPs showed good performances (0.52/4.83% and 0.53/6.24%, respectively), which was in line with the kernel weights. Influenced by these results, we also computed classification models by combining the best-performing views and the views with highest kernel weights.

The corresponding classification results with combined sets of views are shown in the lower block of Table 2. The best set of views was found to be the one with the Alpha, Gamma1, and ERPs (0.56/11.72%). Other significant improvements were achieved with classification models based on Alpha and ERPs, and Gamma1 and ERPs. Even though Beta had a high kernel weight, it did not significantly im-

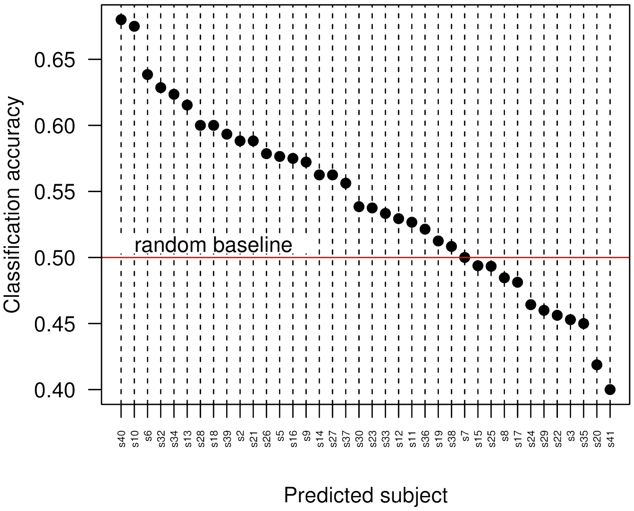

Figure 2: Individual classification accuracy for each of the 38 participants with all seven views based on training on the data of the remaining participants and ordered according to the accuracy. TBRP generalizes for about 70% of the participants which follows the BCI illiteracy.

Table 3: Classification results for a restricted set of participants motivated by the well-known "BCI illiteracy". Bold entries denote that improvements are statistically significant at a level of $\alpha = 0.01$, $p < \alpha$ with correction for multiple testing.

Views	#	Mean accuracy	p-value	Mean improvement
All	26	0.5750	<**0.0001**	15.00%
Al+Ga1	28	0.5641	<**0.0001**	12.82%
Al+E	25	0.5853	<**0.0001**	17.06%
Ga1+E	26	0.5792	<**0.0001**	15.83%
Al+Ga1+Be	25	0.5490	0.0019	9.81%
Al+Ga1+E	28	0.5545	**0.0005**	10.89%

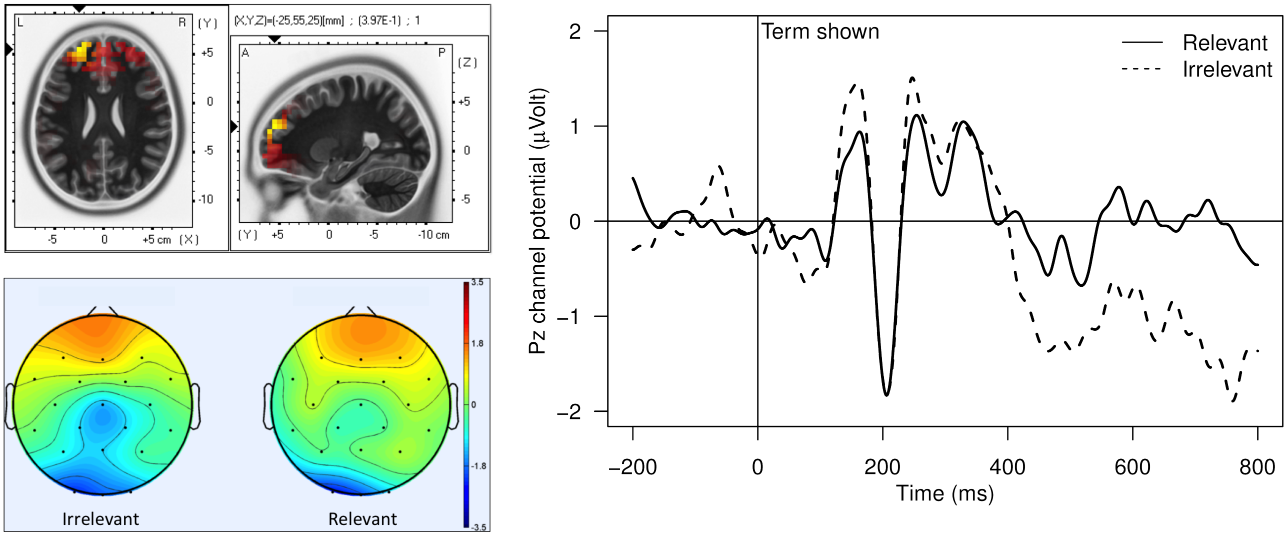

Figure 3: Visualizations of physiological findings based on all 38 participants (details in Section 5.2): (a, top) Localization of the Alpha change associated with relevance mapped to a normalized brain space. Horizontal and sagittal slice show an area of maximum change with peak significant area localized in the Brodmann Area 10, which has been associated with a range of cognitive functions that are important for relevance judgments, such as recognition, semantic processing, memory recall, and intentional planning. (a, bottom) Topography of the event-related potential in the interval between 450 and 747 ms after irrelevant (left) and relevant (right) term onset. Peak significant difference was found in the Pz channel. (b) The corresponding ERP signal at the Pz channel shows the significant difference between relevant and irrelevant after 450 ms, maximizing at 747 ms.

prove the classification accuracy when combined with Alpha and Gamma1. In summary, this suggests that changes in Alpha, Gamma1, and ERPs activities are associated with term-relevance judgments. The physiological findings presented in the next section support these results for Alpha and ERPs.

"BCI illiteracy" analogy. Motivated by the well-known "BCI illiteracy", which means that BCI control does not work for a non-negligible proportion of participants (ca. 15-30%, [39]), we were interested to find out whether a similar effect could be observed in TRPB. In detail, we studied whether we could achieve a better classification accuracy for a specific group of participants. Figure 2 shows the classification accuracy of the classification models using all views; the horizontal line at 0.5 marks the random baseline. 26 participants are above, and 12 participants are below the random baseline. This proportion suggests that capturing the relevance effect is generalizable for 70% of the participants and not generalizable for about 30% of the participants, which follows the BCI illiteracy rate mentioned in [39].

A screening of the EEG signals of the 12 participants did not show a higher noise level, which would explain the worse-than-random prediction accuracy. Given that the learning data for these cases are the other participants (leave-one-participant-out strategy), there may be a group of participants with similar brain signals. There seems, however, to be a group of participants with possibly different brain signals. In order to investigate if we could achieve further classification improvements, we investigated an additional set of classification models for a restricted set of participants. The restricted set was determined via a simple trial-and-error procedure: i.e., we included all participants with an accuracy above the random baseline.

Table 3 shows the corresponding classification results for the restricted set of participants. The results indicate that it is possible to increase the mean prediction accuracy and therefore the mean improvement in all cases; in the best case up to 17% (Alpha and ERPs). Because the simple trial-and-error procedure, these analyses do not provide generalizable results over all participants. This procedure, however, allowed us to demonstrate an analogy to the well-known "BCI illiteracy" effect.

5.2 Physiological Findings

The views that were found most effective for the classification (Alpha and ERPs) were investigated from a physiological point of view. We present brain mappings of the average Alpha effect across all 38 participants and the topography of the strong ERP effect.

Alpha. We attempted to localize the intracranial source of the Alpha using exact low resolution electromagnetic tomography (eLORETA, [29]). eLORETA is a discrete distributed linear weighted minimum norm inverse solution to the source localization of scalp recorded activity, yielding images with exact localization, at a cost of a low spatial resolution. For each participant, two large 1024 ms segments of relevant versus irrelevant terms were used to calculate the cross spectra across all electrodes resulting in 6000 voxels for both relevant and irrelevant terms for each participant. In order to localize the Alpha change associated with relevance, we used a pairwise log of F-ratio test across voxels using spatial normalization to find a maximally significant source localization (with correction for multiple testing [26]).

Table 4: Results of the high-precision classifier based on all 38 participants and all seven views. The average top 5 terms are shown. The terms which are relevant according to the ground truth are in normal font, irrelevant terms according to the ground truth are in italics. Note that, e.g., the irrelevant term "morse code" in the topic "Iraq war" is predicted. A possible explanation is that brain signals associated with "morse code" being relevant for this topic were detected even tough participants finally decided to judge the term as irrelevant.

Topic	Count		Precision	Recall	Top 5 relevant terms
	all	relevant			
Climate change and global warming	209	111	0.52	0.02	Snowmelt, Elevated CO2, Climate change, *hardware synchronization*, *sightseeing*
Entrepreneurship	199	110	0.69	0.18	business risk, startup company, business creation, *shopping*, *virtual relationships*
Immigration integration	204	105	0.52	0.10	citizenship, ethnic diversity, xenophobia, *arsonist*, *morse code*
Intelligent vehicles	185	109	0.80	0.11	pedestrian tracking, collision sensing, remote driving, radar vision, *arsonist*
Iraq war	208	111	0.63	0.15	Saddam Hussein, US army, Tony Blair, *morse code*, *rock n roll*
Precarious employment	204	106	0.57	0.11	minimum wage, employment regulation, job instability, *virtual relationships*, *video-games*
Mean	202	109	0.62	0.13	

The analysis based on the obtained corrected critical two-sided $F* = .37$ results in an area of 10 voxels, all located in the left frontal lobe, specifically in Brodmann Area 10 and a peak localization at MNI coordinates $(-25, 55, 25)$ with a corrected $p < 0.02$; see Figure 3(a, top). The source localization of the effect on alpha oscillations supports the key role of the frontal lobe. The Brodmann Area 10 has previously been related to recognition [32], memory retrieval [33], and the evaluation of working memory [42].

ERPs. To investigate which components of the ERP contribute most to the model, we analyzed the average difference between relevant and irrelevant terms also in a more traditional manner. Average relevant and irrelevant ERPs were computed for each participant over a minimum of 8 and a maximum of 16 correctly classified epochs in each condition. The main significant areas were observed in the Cz, Pz, C4, and P4 channels, with the peak difference in Pz beginning at 477 ms ($p < .05$) and peaking at 757 ms ($p < .0001$); see Figure 3(a, bottom) and Figure 3(b). The latency and topography of the potential suggest the involvement of a P3-like potential. The high latency and parietal topography coincide with the P3b, thus suggesting that relevance does not affect an early change in orientation, but a later, memory-related effect [30].

6. HIGH-PRECISION TRPB

In a practical information retrieval application that can benefit from relevance prediction, the target is to detect true positive examples of terms that represent user's search intent [34]. In such applications, a classifier that trades recall for the benefit of precision can be used to maximize user experience. In other words, a classifier predicts a term as relevant only if the estimated probability of being relevant is very high, i.e., above a certain threshold (high precision). Obviously, the classifier will miss a lot of true relevant terms (low recall). However, we can take advantage of the fact that brain signals can be captured continuously and with high throughput—compared to implicit signals that require explicit user interaction. As a result, a large number of relevance judgments can be observed in a relatively short time. We demonstrate such a high-precision variant of the TRPB classifier and show that it can construct meaningful sets of terms for unknown topics and unseen participants.

6.1 Prediction Setup

One of the advantages of the Bayesian MKL algorithm introduced in Section 4.1 is that its outcome for an unseen term y is not simply the binary decision to relevant or irrelevant, but the predictive distribution of the term being relevant, i.e., $p(y = \text{relevant} \mid \theta)$ with θ the estimated model parameters. We built a high-precision TRPB variant by predicting a term to be relevant only if the probability was higher than 0.99. We used the same prediction setup as in Section 4.2; the models are based on all 38 participants and all seven views. The learned classification models were used to predict the relevance of the terms for an unseen participant and an unknown topic. We then quantified the predicted relevant terms per topic over all unseen participants, which let us to compute the top relevant terms per topic for an average unseen participant.

6.2 Results

The results of the high-precision classifier in predicting relevant terms are shown in Table 4. For each of the six topics, we show the number of observations used in the prediction, precision, and recall achieved by the high-precision classifier, and the terms predicted relevant by the classifier.

While the overall classification problem is still hard, the high-precision classifier achieves a mean precision of 0.62 with an improvement of 25% from the baseline while still sustaining feasible recall of 0.13. However, there are differences in precision across the topics ranging from 0.52 up to 0.8. This suggests that some terms in some topics may have been more difficult for the participants than others.

For example, for the “Entrepreneurship” topic, the classifier was used to classify 199 samples, of which 110 were relevant and the rest were irrelevant. The high-precision classifier reconstructed 29 terms from these samples, of which 20 were relevant and 9 irrelevant, and achieved a precision of 0.69 and recall of 0.18. The top five terms for this topic were “business risk”, “startup company”, “business creation”, “shopping”, and “virtual relationships”. While “shopping” and “virtual relationships” were not relevant for the topic in the strict sense (in the ground truth), they were still predicted relevant by the high-precision classifier. One may argue that these terms are still somewhat relevant for the topic. Similar is the effect of the classifier picking a term that was classified relevant, but assessed irrelevant in the ground truth, is for example the term “Morse code” for the topic “Iraq war” or the term “virtual relationships” for the topic “Precarious employment”. This suggests that the classifier can possibly detect the correct brain pattern of a participant first thinking that the term may be relevant, even when the participant still ends up assessing it irrelevant.

7. DISCUSSION AND CONCLUSIONS

In essence, relevance judgments happen in the brain and therefore the most intriguing way to predict relevance is to directly use the brain signals. Brain signals also have advantages over the more conventional sources of user signals from a practical IR point of view. The recording of the relevance judgments do not require any explicit user interaction, such as user actively clicking on items. The signals can be captured with higher throughput than from explicit user interaction signals. Most practical information retrieval systems assume the interface between the content and the user to be based on user’s expression of the information need using a term representation. Therefore, in order to operationalize brain signals as a part of a real IR system, a central challenge is to predict the relevance of terms based on the brain signals.

We showed that term-relevance prediction using only brain signals captured via EEG is possible. The classification results showed significantly better performances than the random baseline. As a practical application of TRPB, we demonstrated a high-precision relevance predictor and showed that it can construct meaningful sets of terms for unknown topics and unseen participants. To our knowledge, this is the first work utilizing only brain signals as a source for relevance in an IR scenario. Our approach does not require users to explicitly memorize an artificial pattern or a pre-seen object as in BCI or mind-reading research. Moreover, our approach is based on well-established methods, both for the EEG processing and the prediction task, and we were able to support the classification results with physiological findings. The localized brain areas have previously been associated with cognitive functions important for relevance judgments.

While our results show significant improvements, we see several future research directions in order to utilize TRPB as a part of a real IR system. First, our experimental design is balanced between relevant and irrelevant terms in order to ensures that we measure signals and effects related to relevance judgments. In a real IR setting, however, it is likely that the two classes are imbalanced with the majority of the terms being irrelevant. Experiments with more realistic data and larger amount of observations are needed to show how our results generalize to such scenarios. Second, our classification results already generalize over unseen participants, but more sophisticated EEG processing steps and advanced detection methods are needed to automatically cope with the detected “BCI illiteracy” analogy. Third, an obvious next step is to use the predictions as relevance feedback and to quantify the effectiveness of EEG-based relevance feedback as a part of a real interactive IR system. Fourth, we recognize a need for studies that could more specifically reveal the areas of the brain that are activated when users conduct relevance judgements. This could help to reveal the plurality of different mental operations potentially associated with relevance and allow to build non-intrusive wearable EEG systems that could rely on a small number of electrodes at specific positions.

In conclusion, our findings open a horizon for adaptive information retrieval systems that can detect relevance directly from brain signals without requiring users to engage with any particular interaction technique or user interface. With the current trend of wireless, light weight, and portable EEG sensors, our findings can enable systems, which analyze relevance as it happens as a part of our everyday information seeking activities.

8. ACKNOWLEDGMENTS

This work has been partly supported by the Academy of Finland (Multivire, 255725; and the Finnish Centre of Excellence in Computational Inference Research COIN, 251170), Re:Know funded by TEKES, and MindSee (FP7 – ICT; Grant Agreement # 611570).

9. REFERENCES

[1] I. Arapakis, K. Athanasakos, and J. M. Jose. A comparison of general vs personalised affective models for the prediction of topical relevance. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR ’10, pages 371–378, New York, NY, USA, 2010. ACM.

[2] I. Arapakis, I. Konstas, and J. M. Jose. Using facial expressions and peripheral physiological signals as implicit indicators of topical relevance. In *Proceedings of the 17th ACM International Conference on Multimedia*, MM ’09, pages 461–470, New York, NY, USA, 2009. ACM.

[3] H. Berger. Über das Elektrenkephalogramm des Menschen. *Archiv für Psychiatrie und Nervenkrankheiten*, 87(1):527–570, 1929.

[4] F. Crick and C. Koch. A framework for consciousness. *Nature Neuroscience*, 6:119–126, 2003.

[5] L. Farwell and E. Donchin. Talking off the top of your head: Toward a mental prosthesis utilizing event-related brain potentials. *Electroencephalography and Clinical Neurophysiology*, 70(6):510–523, 1988.

[6] B. Figner and R. O. Murphy. Using Skin Conductance in Judgment and Decision Making Research. *A Handbook of process tracing methods for decision research: A critical review and user’s guide*, pages 163–184, 2010.

[7] J. R. Folstein and C. Van Petten. Influence of cognitive control and mismatch on the N2 component of the ERP: A review. *Psychophysiology*, 45(1):152–170, 2008.

[8] G. Gomez-Herrero, W. De Clercq, H. Anwar, O. Kara, K. Egiazarian, S. Van Huffel, and W. Van Paesschen. Automatic removal of ocular artifacts in the EEG without an EOG reference channel. In *Signal Processing Symposium, 2006. NORSIG 2006. Proceedings of the 7th Nordic*, pages 130–133, 2006.

[9] M. Gönen. Bayesian efficient multiple kernel learning. In J. Langford and J. Pineau, editors, *Proceedings of the 29th*

International Conference on Machine Learning (ICML-12), pages 1–8, New York, NY, USA, 2012. ACM.

[10] C. J. Gonsalvez and J. Polich. P300 amplitude is determined by target-to-target interval. *Psychophysiology*, 39(3):388–396, 2002.

[11] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13:723–773, 2012.

[12] T. Gruber, M. M. Müller, and A. Keil. Modulation of induced gamma band responses in a perceptual learning task in the human EEG. *Journal of Cognitive Neuroscience*, 14(5):732–744, July 2002.

[13] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay. Accurately interpreting clickthrough data as implicit feedback. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '05, pages 154–161, New York, NY, USA, 2005. ACM.

[14] D. Kelly and N. J. Belkin. Display time as implicit feedback: Understanding task effects. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '04, pages 377–384, New York, NY, USA, 2004. ACM.

[15] W. Klimesch. EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis. *Brain Research Reviews*, 29(2,3):169–195, 1999.

[16] J. Koenemann and N. J. Belkin. A case for interaction: A study of interactive information retrieval behavior and effectiveness. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '96, pages 205–212, New York, NY, USA, 1996. ACM.

[17] Z. Koldovsky, P. Tichavsky, and E. Oja. Efficient variant of algorithm FastICA for independent component analysis attaining the Cramér-Rao lower bound. *Neural Networks, IEEE Transactions on*, 17(5):1265–1277, 2006.

[18] A. Lagioia, S. Eliez, M. Schneider, J. S. Simons, M. Van der Linden, and M. Debban. Neural correlates of reality monitoring during adolescence. *NeuroImage*, 55(3):1393–1400, 2011.

[19] F. Lotte, M. Congedo, A. Lécuyer, F. Lamarche, and B. Arnaldi. A Review of Classification Algorithms for EEG-based Brain-Computer Interfaces. *Journal of Neural Engineering*, 4(2), June 2007.

[20] S. J. Luck. *An Introduction to the Event-Related Potential Technique (Cognitive Neuroscience)*. MIT Press, 1 edition, 2005.

[21] D. J. McFarland, L. A. Miner, T. M. Vaughan, and J. R. Wolpaw. Mu and beta rhythm topographies during motor imagery and actual movements. *Brain Topography*, 12(3):177–186, 2000.

[22] P. Missonnier, M.-P. Deiber, G. Gold, P. Millet, M. Gex-Fabry Pun, L. Fazio-Costa, P. Giannakopoulos, and V. Ibáñez. Frontal theta event-related synchronization: Comparison of directed attention and working memory load effects. *Journal of Neural Transmission*, 113(10):1477–1486, 2006.

[23] T. M. Mitchell, S. V. Shinkareva, A. Carlson, K. M. Chang, V. L. Malave, R. A. Mason, and M. A. Just. Predicting human brain activity associated with the meanings of nouns. *Science*, 320(4880):1191–1195, 2013.

[24] Y. Moshfeghi and J. M. Jose. An effective implicit relevance feedback technique using affective, physiological and behavioural features. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '13, pages 133–142, New York, NY, USA, 2013. ACM.

[25] Y. Moshfeghi, L. R. Pinto, F. E. Pollick, and J. M. Jose. Understanding Relevance: An fMRI Study. In P. Serdyukov, P. Braslavski, S. Kuznetsov, J. Kamps, S. Rüger, E. Agichtein, I. Segalovich, and E. Yilmaz, editors, *Advances in Information Retrieval*, volume 7814 of *Lecture Notes in Computer Science*, pages 14–25. Springer Berlin Heidelberg, 2013.

[26] T. E. Nichols and A. P. Holmes. Nonparametric permutation tests for functional neuroimaging: A primer with examples. *Human Brain Mapping*, 15(1):1–25, 2002.

[27] M. Pantic and L. J. M. Rothkrantz. Automatic analysis of facial expressions: The state of the art. *IEEE Transactions in Pattern Analysis Machine Intelligence*, 22(12):1424–1445, Dec. 2000.

[28] M. Pantic and L. J. M. Rothkrantz. Toward an affect-sensitive multimodal human-computer interaction. *Proceedings of the IEEE*, 91(9):1370–1390, 2003.

[29] R. D. Pascual-Marqui, C. M. Michel, and D. Lehmann. Low resolution electromagnetic tomography: A new method for localizing electrical activity in the brain. *International Journal of Psychophysiology*, 18(1):49–65, 1994.

[30] J. Polich. Updating p300: An integrative theory of P3a and P3b. *Clinical Neurophysiology*, 118(10):2128–2148, 2007.

[31] A. T. Pope, E. H. Bogart, and D. S. Bartolome. Biocybernetic system evaluates indices of operator engagement in automated task. *Biological Psychology*, 40(1–2):187–195, 1995. EEG in Basic and Applied Settings.

[32] C. Ranganath, M. K. Johnson, and M. Desposito. Prefrontal activity associated with working memory and episodic long-term memory. *Neuropsychologia*, 41(3):378–389, 2003. Functional Neuroimaging of Memory.

[33] M. D. Rugg, P. C. Fletcher, C. D. Frith, R. S. J. Frackowiak, and R. J. Dolan. Differential activation of the prefrontal cortex in successful and unsuccessful memory retrieval. *Brain*, 119:2073–2083, 1996.

[34] T. Ruotsalo, J. Peltonen, M. Eugster, D. Glowacka, K. Konyushkova, K. Athukorala, I. Kosunen, A. Reijonen, P. Myllymäki, G. Jacucci, and S. Kaski. Directing exploratory search with interactive intent modeling. In *Proceedings of the 22nd ACM international conference on Conference on information and knowledge management*, CIKM '13, pages 1759–1764, New York, NY, USA, 2013. ACM.

[35] Y. Saeys, I. Inza, and P. Larrañaga. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19):2507–2517, 2007.

[36] T. Saracevic. Relevance: A review of the literature and a framework for thinking on the notion in information science. part III: Behavior and effects of relevance. *Journal of the American Society for Information Science and Technology*, 58(13):2126–2144, Nov. 2007.

[37] S. Sonnenburg, G. Rätsch, C. Schäfer, and B. Schölkopf. Large scale multiple kernel learning. *Journal of Machine Learning Research*, 7:1531–1565, 2006.

[38] M. M. Spapé and D. J. Serrien. Interregional synchrony of visuomotor tracking: Perturbation effects and individual differences. *Behavioural Brain Research*, 213(2):313–318, 2010.

[39] C. Vidaurre and B. Blankertz. Towards a cure for BCI illiteracy. *Brain Topography*, 23(2):194–198, 2010.

[40] R. W. White and D. Kelly. A study on the effects of personalization and task information on implicit feedback performance. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, CIKM '06, pages 297–306, New York, NY, USA, 2006. ACM.

[41] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan. Brain-computer interfaces for communication and control. *Clinical Neurophysiology*, 113(6):767–791, 2002.

[42] J. X. Zhfang, H.-C. Leung, and M. K. Johnson. Frontal activations associated with accessing and evaluating information in working memory: an fMRI study. *NeuroImage*, 20(3):1531–1539, 2003.